ARL-TR-9067 ● Sep 2020

Human–Autonomy Teaming: Using Latent Semantic Analysis for Assessing Team Cohesion from Crew Communication

by Daniel E Forster, Sean Michael McGhee,
Brandon S Perelman, Catherine Neubauer, Kristin E Schaefer,
and Andrea Krausman

NOTICES

Disclaimers

ARL-TR-9067 ● SEP 2020

Human–Autonomy Teaming: Using Latent Semantic Analysis for Assessing Team Cohesion from Crew Communication

Daniel E Forster, Brandon S Perelman, Catherine Neubauer,
Kristin E Schaefer, and Andrea Krausman
Human Research and Engineering Directorate, CCDC Army Research Laboratory

Sean Michael McGhee
DCS Corporation

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* September 2020	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** July 2019–September 2020

4. TITLE AND SUBTITLE Human–Autonomy Teaming: Using Latent Semantic Analysis for Assessing Team Cohesion from Crew Communication	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Daniel E Forster, Sean Michael McGhee, Brandon S Perelman, Catherine Neubauer, Kristin E Schaefer, and Andrea Krausman	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) CCDC Army Research Laboratory ATTN: FCDD-RLH-FA Aberdeen Proving Ground, MD 21005	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-9067
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

ORCID ID(s): Daniel E Forster, 0000-0001-8351-2009; Catherine Neubauer, 0000-0002-6686-3576; Brandon S Perelman, 0000-0001-6558-0286; Kristin E Schaefer, 0000-0002-1342-3446; Andrea Krausman 0000-0003-1955-8867

14. ABSTRACT

Team cohesion is a critical component of team function and effectiveness. With respect to military operations, there is a need to assess team cohesion in real time, over the course of rapidly changing circumstances, as unobtrusively as possible. With the emphasis on integrating autonomous agents as team members, the need to understand team cohesion is even more critical. The main goal of this research was to evaluate the effectiveness of using latent semantic analysis (LSA) to quantify team cohesion from recorded verbal crew communication during a simulation study of human–autonomy teaming operations. In the present work, a six-person crew interacted with a manned and two unmanned robotic vehicles using different transparency displays during simulated cordon and search missions. Result from these analyses showed that additional assessment methods, such as LSA, can help explain the real-time effects of human–autonomy team dynamics on team cohesion.

15. SUBJECT TERMS

verbal communication, team communication, speech-to-text transcription, Dragon software

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Daniel Forster
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	45	**19b. TELEPHONE NUMBER (Include area code)** (410) 278-1280

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

Summary

Team cohesion is defined as an emergent property that characterizes a shared bond that motivates team members to stay and work together in pursuit of a common goal.[1,2] While team cohesion has been studied for decades, there remains a need to address the dynamic, emergent nature of team cohesion in mixed human–autonomy teams. Self-report measures are frequently used; however, self-report scales have limitations that make them less appropriate for assessing human–autonomy teams performing complex tasks in real time. Therefore, unobtrusive assessment methods are needed to capture the dynamic nature of team cohesion that should be capable of assessing the construct in near real time, over the course of a mission (and varying timescales). With such methods, researchers can more fully understand the cohesion–performance relationship and identify approaches for strengthening human–autonomy team cohesion when it becomes unstable.

This report explores the potential of latent semantic analysis (LSA) of crew communications as a means of deriving metrics to maximize team performance within military human–autonomy teams. This research is critical to advancing team effectiveness because communication is the mechanism by which shared mental models are built, and natural language allows people to quickly build new models (and update old ones) as needed. Communication content provides rich insights into team functions and processes, and it is methodologically easy to collect unobtrusively. LSA, the linguistic analysis method on which this report is focused, quantifies the semantic structure of documents from observed word co-occurrences, which researchers have used to model properties of communication among teams (e.g., Dong[3]). For this work, LSA was used to analyze crew communications during a simulated cordon and search mission, while they interacted with an autonomy-enabled manned vehicle and two weaponized robotic combat vehicles. Results showed that greater similarity between speakers' utterances in semantic space (constructed through LSA) positively correlated with team members' levels of self-reported belongingness and negatively correlated with the amount of time team members spent facing task difficulties, suggesting that methods such as LSA complement self-report measures and help explain human–autonomy team effectiveness.

---

[1] Mathieu J, Maynard MT, Rapp T, Gilson L. Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. Journal of Management. 2008;34(3):410–476.

[2] Casey-Campbell M, Martens ML. Sticking it all together: A critical assessment of the group cohesion-performance literature. International Journal of Management Reviews. 2009;11:223–246.

[3] Dong A. The latent semantic approach to studying design team communication. Design Studies. 2005;26(5):445–461.

1. Introduction

The US Army Futures Command (AFC) established Cross-Functional Teams (CFTs) to address the needs of the future force in order to maintain superiority on the battlefield. The Next-Generation Combat Vehicle (NGCV) CFT's goal is to address the need for new vehicle platforms with enhanced technological capabilities and the ability to dominate in close combat within Multi-Domain Operations (MDO)*. To reach this goal, the NGCV CFT proposes to integrate autonomous systems, advanced sensors, and Soldiers into heterogeneous teams that are more effective together than any entity acting alone. This concept involves mixed human–autonomy teams consisting of optionally manned combat vehicles (MCVs) and multiple autonomy-enabled weaponized robotic combat vehicles (RCVs) functioning as a connected warfighting element. Equipping these RCVs with advanced sensors and autonomous capabilities may enhance Soldier survivability by minimizing exposure to direct threats, while providing additional "eyes" and "ears" distributed across the battlefield to increase situation awareness and reduce the time required to make well-informed decisions. Essentially, this creates a paradigm shift as we move from viewing autonomous agents as "tools" used by Soldiers to accomplish a mission, to viewing them as full-fledged members of the team. However, it is important to recognize that autonomous systems behave, reason, and communicate differently than humans. This poses unique challenges for effective teaming between Soldiers and autonomous systems that are critical to team processes, such as developing effective team communication, as they contribute to the emergence of team constructs, such as team cohesion.

While the literature on the cognitive underpinnings of effective *human* team performance is broad and rich (DeChurch and Mesmer-Magnus 2010), to apply these findings directly to human–autonomy teams requires extensive research on the inherent differences between human teams and mixed human–autonomy teams. In doing so, it is necessary to understand how these differences affect the ability of human and autonomous team members to coordinate, communicate, and understand one another, and how these differences fundamentally change teamwork dynamics. To address these challenges, the US Army Combat Capabilities Development Command (CCDC) Army Research Laboratory (ARL) created the Human–Autonomy Teaming Essential Research Program (HAT ERP), of which one primary goal is to address the challenges associated with integrating humans and autonomous systems into teams that work cooperatively and maintain

---

*The US Army's MDO 2028 concept incorporates three tenets: calibrated force posture, multi-domain formations, and convergence for overmatch in contested operational environments that are increasingly lethal, complex, and challenged (TRADOC 2018).

appropriate levels of team trust and cohesion within the complex NGCV environment. As a result, there exists a need to generate new metrics and methods that are appropriate for assessing the performance of human–autonomy teams as well as emergent team constructs, such as trust and cohesion, that are essential for effective teamwork within the complex NGCV context. There are many factors that can influence the various team constructs mentioned previously; however, communication is the very mechanism by which people, especially in teams, build shared representations of how to coordinate their actions to navigate various environments (Sperber and Wilson 1986). Therefore, this report focuses on a novel approach for assessing team cohesion through analyzing team communication with latent semantic analysis (LSA), which will contribute to a better understanding of how humans and autonomy interact with one another, and how cohesion emerges and is maintained in human–autonomy teams.

1.1 Human–Autonomy Teaming: Team Cohesion

Research suggests that comprehension of team members' actions and intentions, as well as their understanding of the goals and general reasoning of the task, can impact how an individual contributes to the team (Mathieu et al. 2000). Further, studies repeatedly show that mutual understanding of actions and intentions facilitates the development of team cohesion, which can be construed as an emergent state characterized by team members' shared commitment to achieving a common goal (Mathieu et al. 2008). Team cohesion is thus influenced by various constructs that keep members motivated to work and stay together (Casey-Campbell and Martens 2009). The team properties that facilitate the emergence of cohesion are especially important for military teams because cohesion has a direct impact on team behaviors and performance (Beal et al. 2003). For instance, Oliver et al. (1999) used measures of cohesion to predict successful team performance; a finding reinforced by Ahronson and Cameron's (2007) observation that cohesive military teams report being more effective in combat situations. Furthermore, the feedback effects of cohesion on teams can vary depending on the team's context, the antecedent factors that shape team members' interactions, team members' interactions directed toward accomplishing their tasks, and the outcomes of these interactions (Marks et al. 2001; Mathieu et al. 2008). Team cohesion is critical for team functioning and effectiveness, and several meta-analyses have emphasized the necessity of cohesion for tasks that are complex with a high degree of interdependence, such as those performed by medical, military, emergency response, aviation, and other teams operating in extreme environments (Santoro et al. 2015). Cohesion is also positively associated with outcomes such as job satisfaction, psychological well-being, team viability, and collective efficacy, to name a few (Ahronson and Cameron 2007; Grossman 2014). Some scholars have even found evidence that team cohesion is a critical factor in emotional resilience following periods of stress (Neubauer et al. 2016). While on the surface it appears that increased cohesion would be a consistent outcome of interest, researchers have found that high levels of cohesion can also be destructive to a team because it can cause a so-called "groupthink" effect, whereby individuals think less critically about information shared among the team due to conformity motivations (Janis 1982; Gigone and Hastie 1993; Gruenfeld et al. 1996). For team members to engage with team processes in a manner that produces adaptive levels of cohesion, proper measurement of the construct must be paired with more rigorous attempts to disentangle "team cohesion" from its antecedents and consequents.

When measuring team cohesion, researchers predominantly use subjective self-report scales to assess each member's perceptions of cohesion. In fact, some scholars argue that cohesion ought to be measured with self-report scales, as observer reports of team cohesion may, in fact, be assessing a different construct entirely (Abrams and Rosenthal-von der Putten 2020; see Kozlowski [2015] for a different view). To understand why measures of team cohesion are important for understanding performance, it is useful to consider how self-reported cohesion relates to other team constructs. For instance, teams who report higher levels of cohesion also report greater satisfaction with their team (Curtis and Miller 1986; Dobbins and Zaccaro 1986; Ahronson and Cameron 2007), more effective and timely communication (Onağ and Tepeci 2014; Kim et al. 2016), increased trust (Mach et al. 2010), and an increase in helping behaviors (Ng and Van Dyne 2005; Liang et al. 2015). However, this wealth of constructs related to team cohesion has been obtained at the price of not understanding how these constructs change and interact at various stages of the team's lifespan. Consider, for instance, the limitations imposed by relying on self-reported measures of cohesion: 1) subjective responses are highly vulnerable to various, often uncontrollable, sources of bias, and 2) measures are typically performed at a single time point, resulting in responses reflecting a person's summary of their highly dynamic experience. Further, as we consider assessments of cohesion in teams in which autonomy is considered a teammate, there are constraints on the autonomous agent's ability to provide meaningful self-report data, which, to be appropriately compared to human responses, ought to be generated from a comparable causal mechanisms as human responses (Borsboom 2006). Therefore, novel methods that go beyond looking at traditional subjective measures for assessing team cohesion are needed to best understand this dynamic within human–autonomy teams.

To evaluate cohesion without relying on self-report measures, some scholars have investigated the use of verbal communication between crew members as an indicator of team cohesion (e.g., Hill et al. [2002]). Communication metrics of cohesion are particularly appealing for our purposes, as they have demonstrated utility for human–human and human–agent interactions (Hill et al. 2002; Onağ and Tepeci 2014; Kim et al. 2016 Dhall 2019). Further, the relative ease in which communication data can be obtained in the laboratory or field, and the insights it provides with respect to understanding cognitions, emotions, and emergent team processes and states (Kiekel et al. 2002; Foltz and Martin 2008) make it especially valuable as a metric. Evaluating team cohesion using communication metrics can be accomplished in a variety of ways by assessing the dialogue expressed between teammates (e.g., what kind of language team members use with each other [Gorman et al. 2003; Dong 2005]), the spectral content of the voice (Neubauer et al. 2016, 2017), nonverbal features like body posture and facial expressions (Behoora and Tucker 2015; Neubauer et al. 2016; Dhall 2019), and even technologically mediated communication (e.g., text chat, button presses [Whitman et al. 2005; Ehsan et al. 2008]). For example, Neubauer and colleagues (2016, 2017) found that team members who used language relating to "cognitive processing" and problem solving exhibited better cohesion and resilience following a stressful dyadic task compared to teams who used language relating to "emotion or affect". The implication here was that overly emotional language may interfere with an individual or team's ability to effectively communicate necessary information. For communication to be effective during a time-limited, stressful task, team members must relay only critical information, while also avoiding extraneous irrelevant detail, which may be distracting.

Several theoretical and methodological contributions to communication analysis have yielded promising applications for teams. For instance, scholars have contributed to ever-growing dictionaries, in which words are viewed as weighted indicators of cognitive, emotional, and social processes (e.g., Pennebaker et al. [2001]). Others have, instead, relied upon models in which the meaning and relevance of communication is built from the structure of the communication itself; by using co-occurrences of words and other relevant information, one can create a high-dimensional semantic space in which observed content is used to infer unobserved topics of communication (e.g., Dumais [2004]). One such method is latent semantic analysis (LSA), which is the focus of the current investigation. LSA has been used by scholars as a basis for modeling approaches that produce human-level performance on low-level linguistic tasks, such as correctly inferring word meanings in a multiple-choice testing format, as well as higher-level linguistic tasks, such as annotating text from human communications in a mixed human–autonomy crew scenario (Martin and Foltz 2004; Foltz et al. 2006). Given the attention LSA has received as a viable approach for automatically analyzing communication, we focused on using LSA to derive metrics of team cohesion for human–autonomy teams.

1.2 Latent Semantic Analysis

LSA is a method for analyzing a corpus of data, typically comprising written text or transcribed speech nested within several documents (for reviews, see Landauer et al. [1998] and Dong [2005]). By examining a term-document matrix, in which rows represent unique terms (e.g., words), columns represent unique documents (e.g., utterances), and cells represent the number of times a term appears in a document, LSA can infer term meanings based on co-occurrences with other words. To do so, LSA relies on a model in which words and documents are related due to their relevance to unobserved (i.e., latent) topics; that is, topics dictate the decisions we make regarding which words we use, so any observed word clustering ought to be, at least to some extent, indicative of such a topic. The output of such an analysis is a dimensionality reduction of the raw term-document matrix, this time with cells representing the semantic relevance of each term with each document. In the case of team communication, such an analysis could be conducted to examine the semantic features underlying speakers' utterances, which could be further analyzed to assess coherence between speakers and the efficiency (i.e., information density) with which speakers communicate.

Metrics derived from LSA have been used to predict positive social dynamics, such as greater interpersonal attentiveness (Babcock et al. 2014) and greater team performance (Gorman et al. 2003; Dong et al. 2004; Martin and Foltz 2004). For example, Gorman et al. (2003) used LSA to evaluate communication in three-person teams that operated simulated Predator unmanned aerial vehicles. The authors found that team communication density was related to performance, and that LSA was useful in distinguishing high- and low-performing teams, which was based on the time and resources used to complete the mission. In more recent work, Gorman et al. (2013) found that the length of utterances in semantic space could be used to differentiate between experienced and less-experienced teams during submarine training. In addition, their LSA-derived metric of communication synchrony correlated with a metric of team neurophysiological synchrony, providing some convergent evidence that linguistic and neural synchrony reflect the same underlying team processes, as well as suggesting that future work might further develop a paradigm for understanding team performance through a framework involving team communication, team neurophysiology, and other domains beyond communication content.

1.3 Current Work

The purpose of this work was to evaluate LSA's potential as a method for quantifying cohesion in human–autonomy teams from their communications data. Although some previous research suggests that LSA may be viable for quantifying team cohesion (Dong 2005), it is necessary to understand how this methodology applies to future human–autonomy teams within Army operations. To address this gap, communications data were collected during a simulation experiment conducted in support of the NGCV modernization priority.

2. Methods

These analyses are part of a larger research study aimed at assessing the impact of transparency aids on different human–autonomy teams' goals, intentions, reasoning, and actions. A full review of the study goals is available in ARL-TN-1003 (Perelman et al. 2020a) and ARL-TR-9002 (Perelman et al. 2020b). The order of events and schedule for the larger study are located in the Appendix. This report focuses on identifying potential metrics of team cohesion through LSA.

2.1 Participants

Six crew-members from the Soldier, Operator, Maintainer, Test, and Evaluation (SOMTE) unit at Aberdeen Proving Ground participated in the experiment. The same crew members participated in all mission runs and maintained the same roles throughout the experiment.

2.2 Apparatus and Facilities

The CCDC Army Research Laboratory Information for Mixed Squads (INFORMS) laboratory contains a stationary 7-person simulated crew station organized to match the setup of the future NGCV command and control vehicle (Fig. 1). During the experiment, participants were assigned to a crew position or role within the simulated MCV to control all vehicles in the NGCV squad. The MCV frame contained seven crew stations; one crew station at the front of the frame, and three additional pairs of crew stations. Participants completed the experiment using the three pairs of crew stations. Each dyad controlled one of the simulated vehicles, either the MCV or one of the two RCVs. The seventh crew station was used by a member of the experimental team to provide goal locations to the squad.

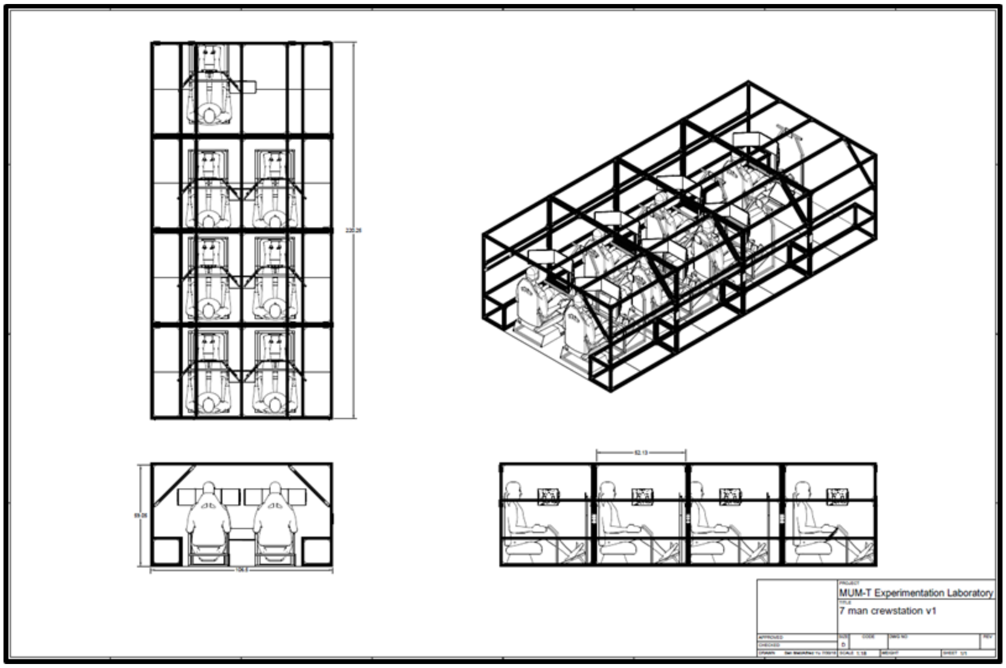

Fig. 1 MCV layout in the INFORMS lab. Each row contains a dyad, primarily tasked with controlling (from front to back) the MCV itself and the two RCVs. The front-most crew station seats a section-level commander. Source: ARL-TN-1003 (Perelman et al. 2020a).

For this experiment, each computer station used Autonomous Navigation Virtual Environment Laboratory (ANVEL; Quantum Signal) simulation for the virtual environment and the Warfighter Machine Interface (WMI; DCS Corporation) user interface for interaction with a simulated MCV and two simulated RCVs. Each crew station was designed with three touchscreen displays and human interface devices, such as an instrumented steering yoke and pedals (Fig. 2).

Fig. 2 Example crew station with three touchscreen displays and vehicle controls that allow participants to interact with the simulated manned and robotic vehicles

2.3 Task

During the simulated missions, the crew was responsible for planning and executing movement of the MCV and two RCVs to set objectives while avoiding areas of the environment containing the types of dynamic threats expected in MDO and maintaining 360° situation awareness to identify targets in the environment. Figure 3 displays an example map of the environment, which participants used to plan routes for manned and robotic vehicles.

Fig. 3 Example map layout for planning routes for manned and robotic vehicles

2.4 Design

The experiment was a two WMI (Baseline WMI vs. Transparent WMI) × two Path Planning (Autonomy vs. Operator) within-team factorial design, with two runs in each condition, for a total of eight runs. For this analysis, only one of each condition (total four runs) were analyzed due to incomplete transcription data. The differences between these conditions are described in more detail:

- *Baseline WMI (Runs 3 and 4):* The Baseline WMI is the version received from CCDC Ground Vehicle Systems Center in 2018. This version provided crew members the ability to teleoperate or plan routes to autonomously drive the vehicle (move) through waypoint navigation, and transmit mission-relevant data using a map-based common operating picture (communicate). For more information on this version of the WMI, see Perelman et al. (2020a).

- *Transparent WMI (Runs 1 and 2):* The Transparent WMI was the Baseline WMI software plus additional hardware and software designed to improve transparency in the vehicles' actions, intentions, goals, and general reasoning. Software modifications included changes to the route planning functionality (Transparent Route Planner; Fig. 4) to enable supervisory control versus deliberate waypoint-by-waypoint planning, and a visualization (Comparator Display; Fig. 5) to enable commanders to compare multiple routes simultaneously and select the best based on mission, enemy, terrain, troops available, time, and civil considerations (Perelman et al. 2020b). Hardware modifications included a multimodal (e.g., auditory and tactile) cueing system that improved crew members' local situation awareness by alerting them to the presence of threats in the environment and vehicle status updates (see Fig. 6; Chhan et al. [2020]).

Fig. 4 Transparent Route Planner (for supervisory control) explains the reasoning behind the autonomy's planning decisions

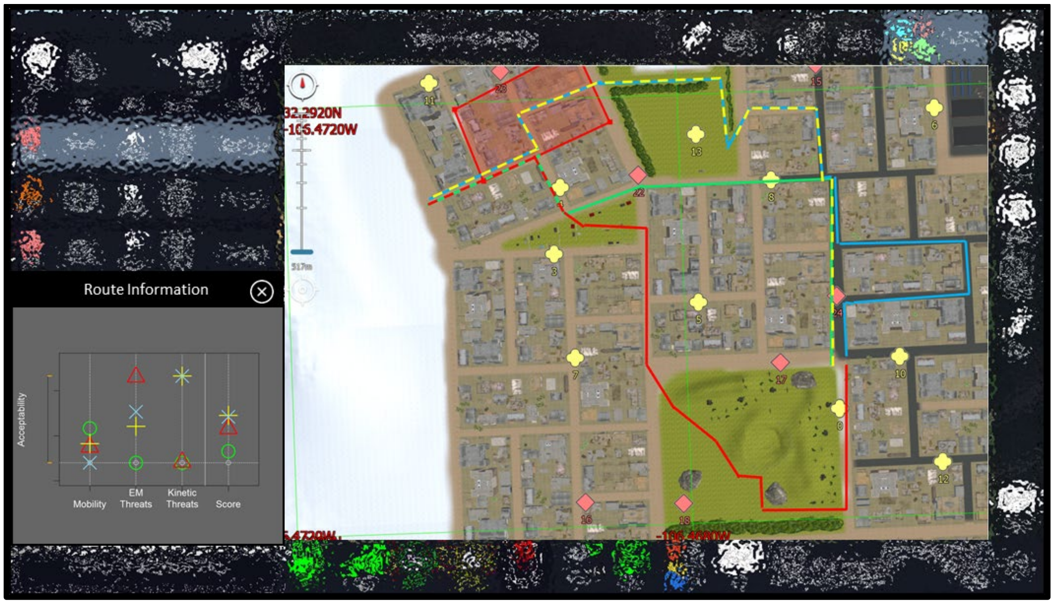

Fig. 5 Comparator Display enables direct comparisons of alternative route plans

Fig. 6 A tactile belt is used as part of the Transparent WMI condition to provide cuing to crew members whenever they encountered an environmental threat. Diagram (bottom left) depicts tactor position in the belt.

- *Autonomy Planning (Runs 2 and 4):* To ensure that vehicle crews would encounter at least a certain number of dynamic threats and targets in the environment, the section completed what was initially noted in the experimental design as constrained route reconnaissance missions. During the constrained runs, each vehicle was provided with its own predefined route to a series of objective rally points and was asked to avoid threats and mark targets along the way. These routes were initially generated using the Transparent Route Planner autonomous route planning software, so therefore are described here as autonomy-planned routes

- *Operator Planning (Runs 1 and 3):* Vehicle crews also completed comparatively open cordon and search missions, during which they were provided a series of locations to search (indicated by MIL-STD 2525D *Occupy* mission graphics [DOD 1994]) as a section. This mission required the NGCV section to coordinate their search patterns and was designed to encourage coordination via radio communication. This condition is described herein as operator-planned routes since the vehicle crews were responsible for planning the vehicles' mobility.

2.5 Voice Communication Recording and Transcription

During the experiment, participants communicated with one another using the commercial off-the-shelf (COTS) HyperX Cloud Pro Gaming Headsets with a microphone. Microphone amplifiers were used for each crew member to minimize signal loss. Audio cables were run from each crew position to a 16-channel mixer and recorded as separate .wav files using a laptop running Adobe Audition software.

Dragon Naturally Speaking 15.1, a COTS software from Nuance, and Simple Direct Media Layer (SDL 2) were deployed on all workstations. Dragon captured speech and rendered it to text and SDL captured button presses on the console device on each station. A client application that interfaced with these packages was also deployed to each workstation and used Dragon to capture headset microphone input whenever one of two buttons on the console were pressed. One button was for all open communication with all six team members, while the second button allowed for only communication with the direct team member.

Although Dragon was implemented to ease the process of converting speech to text, analyses revealed that the real-time transcriptions failed to capture large chunks of speech. This appeared to be due, in part, to limitations with real-time analysis, because the same software performed somewhat better when attempting to transcribe communications after audio recording was completed. Unfortunately, even the results of post-hoc speech-to-text transcriptions were far from accurate—when compared to manually transcribed audio files, Dragon transcriptions were between 12%–67%, with only three transcriptions exceeding 50% accuracy. Therefore, the analyses reported here rely solely on the manually transcribed data, as this would provide the best estimates for how communication metrics would relate to team cohesion and performance.

2.6 LSA Method

LSA was conducted on the transcribed communication data using the *lsa* package (Wild 2020) in R version 3.6.1 (R Core Team 2019). To create a semantic space under which the communication transcripts were analyzed, we first removed all so-called "stop words" from the transcripts, which represent function words that contain no semantic information (e.g., "the", "yes", "there", and so on) and would therefore be uninformative regarding the topics underlying the semantic space (Silva and Ribeiro 2003). After reducing the transcripts to only semantically relevant words, a term-document matrix was created with each utterance representing a unique document and each word representing a term in the document, resulting in a sparse matrix in which each cell represents the term frequency (i.e., the number of times a word appeared in an utterance). Subsequently, we transformed the matrix using a weighting function, which gives less weight to those words that are common across all utterances, and more weight to those words that are specific to fewer utterances—referred to as the term frequency-inverse document frequency (Robertson 2004).

In the analyses presented here, we created these matrices for each dyad within the team—that is, communications between each MCV, RCV1, and RCV2 pair were used to create a distinct latent semantic space for comparison. Further, we created distinct matrices for each of the four runs (each of which represents a different experimental condition), so each latent semantic space represents a unique combination of dyad (MCV, RCV1, and RCV2) and run (1–4; see Table 1 for which conditions correspond to which runs), allowing us to examine how communication features extracted with LSA correspond to self-reported cohesion and performance.

Descriptive information for each dyad's corpus in each run is reported in Table 1 The largest corpus consisted of 303 utterances and 624 words (RCV1 dyad when using autonomous planning with Transparent WMI), whereas the smallest corpus consisted of 60 utterances and 90 words (RCV2 dyad when using autonomous planning with Baseline WMI). Although researchers argue that extracting approximately 300 topics (or more) results in best performance (e.g., Dong [2005]), these recommendations are often considered in large corpora. Ultimately, the number of topics extracted using LSA cannot exceed the number of documents; therefore, to simplify these analyses, only 25 topics were extracted for each corpus.

Two metrics were computed within each corpora's semantic space: communication density and utterance coherence between speakers. The following paragraphs outline how each of these metrics were computed.

Table 1 Descriptive information about crew communication

Dyad	Run	Mobility	WMI	Utterances	Words
MCV	1	Operator	Transparent	137	161
	2	Autonomy	Transparent	177	204
	3	Operator	Baseline	165	176
	4	Autonomy	Baseline	80	91
RCV1	1	Operator	Transparent	141	193
	2	Autonomy	Transparent	303	624
	3	Operator	Baseline	126	225
	4	Autonomy	Baseline	102	177
RCV2	1	Operator	Transparent	165	193
	2	Autonomy	Transparent	155	189
	3	Operator	Baseline	125	152
	4	Autonomy	Baseline	60	90
Total	1	Operator	Transparent	443	547
	2	Autonomy	Transparent	635	1,017
	3	Operator	Baseline	393	553
	4	Autonomy	Baseline	242	358

Note: Word counts exclude the use of stop words (i.e., semantically irrelevant words, such as "the", "a", and "yeah").

2.6.1 Communication Density

Communication density represents the rate of information flow across utterances, or the degree to which words relate to each utterance in semantic space—the more semantically relevant words for any given utterance would indicate that dyads are communicating more efficiently. Computationally, the semantic information within an utterance represents the sum of vector lengths within that utterance in semantic space. Therefore, the average communication density across utterances is computed using the total sum of the semantic space (i.e., the total of all semantically relevant information in the conversation) divided by the number of utterances.

2.6.2 Between-Speaker Coherence

The coherence between utterances is computed via the cosine similarity between the utterances in semantic space. To break this down, LSA creates a semantic space in which words and utterances are associated due to their relevance to an underlying topic—these topics permeate words and utterances, the consequences of which allow words unspoken in an utterance to nevertheless be related to that utterance, by virtue of the fact that they lie on the same topic dimension. Taking this a step further, an utterance's occupation of semantic space can be correlated with another utterance's occupation of the same space, which is achieved by computing the cosine similarity between utterances. Similarly, words can be associated with each other by virtue of their similar occupation of semantic space, as well as words with documents. Using this method, we computed the average cosine similarity exclusively between speakers' utterances. Gorman et al. (2013) noted that they used a 36-utterance moving window to compute coherence between speakers. By doing so, they were acknowledging that early utterances may be unrelated to later utterances for reasons having nothing to do with dyads speaking coherently. Rather than using a moving window, we simply computed the average cosine similarity between dyads across all utterances.

2.7 Measures

While a number of measures were included in the full study, only those being analyzed in this report are described here.

2.7.1 Perceived Cohesion

Self-reported cohesion was measured using Chin et al.'s (1999) small group adaptation of the six-item Perceived Cohesion Scale (Bollen and Hoyle 1990). This scale assesses two dimensions of cohesion: Belongingness (e.g., "I feel that I belong to this group.") and Morale (e.g., "I am happy to be part of this group."). Participants responded to each item on a 7-point ordinal scale (1 = *Strongly Disagree*, 7 = *Strongly Agree*). Assessments of Cronbach's α at each time point exhibited good internal consistency (Table 2). We also assessed average interrater agreement across all six items using James et al.'s (1984, 1993) measure of within-group interrater agreement, r_{wg}, specifically the extension for multiple items, $r_{wg.j}$. Although researchers use r_{wg} values, which range from 0 (no agreement) to 1 (perfect agreement), and recommended thresholds (e.g., $r_{wg} \geq 0.70$) to justify intragroup data aggregation, these practices remain controversial (Woehr et al. 2015); therefore, we report this merely for its descriptive value. Ultimately, we computed a dyad-level measure of perceived cohesion by averaging across all item responses within a dyad, separately for each run.

Table 2 Cohesion internal consistency, dyadic agreement, and inter-construct correlation across conditions

Condition	Mobility planning	WMI	Morale	Belongingness	*r*
			Cronbach's α ($r_{wg.j}$)		
1	Operator	Transparent	0.967 (*0.768*)	1 (*0.833*)	*0.964*
2	Autonomy	Transparent	0.797 (*0.5*)	0.73 (*0.7*)	*0.945*
3	Operator	Baseline	1 (*0.833*)	1 (*0.25*)	*0.985*
4	Autonomy	Baseline	0.963 (*0.693*)	0.984 (*0.482*)	*0.907*

2.7.2 Mobility Challenge Duration

Our first performance measure captures vehicle crews' ability to mitigate difficulties with maneuvering the vehicle during the mission. We measured the duration of mobility challenges, which we defined as any instance in which the vehicle is operating autonomously and moving at or below 1 m/s for 5 s. Importantly, this variable does not count seconds of mobility challenge until *after* the 5-s threshold has passed, meaning that a value of 1 would indicate that the vehicle had spent 6 s traveling at or under 1 m/s. Lower values indicate the crew was able to mitigate mobility challenges more quickly and is therefore indicative of better performance. The final measure represents the average time, during each run, that each dyad spent dealing with mobility challenges.

2.7.3 Threat Response Distance

The second performance measure represents the distance from the threat at which the vehicle indicated a threat response, such as slowing or stopping the vehicle. Higher values indicate a greater distance from the threat, which is indicative of better performance. The final measure represents the average response distance from all threats during a run, separately for each dyad.

2.7.4 Average Threat Proximity

Although threat response distance gives a good sense of dyads' initial threat response, it cannot directly tell how efficiently they avoided threats over the course of the entire run (for example, maintained awareness over the map screen to avoid threats that had already spawned during the mission). To capture this construct, we used an accumulating value representing the threat proximity, p, that accumulated at each time step, $r - d$, where r is the threat's radius of effect (40 m in all cases) and d is the distance between the vehicle and the threat's location. The final measure represents the average of this accumulating value across all threats within each run-dyad combination, with higher values being indicative of worse performance.

2.8 Conceptual Model

Before introducing the results, it is important to clarify the analytical models of team cohesion and how they are used to address relevant questions using these types of data. Performance and communication metrics are summary scores of events that occurred during a complex mission (i.e., linguistic summaries from communication metrics and performance summaries from behavioral metrics). Similarly, self-reports from participants were recollections regarding team dynamics during the missions (i.e., perceived cohesion).

Even though cohesion is an emergent property of groups that results from members developing relevant and timely interdependent behaviors over time, the measure of perceived cohesion used in these analyses represents participants' summaries of these dynamic processes, which were administered after the relevant behavioral actions had already taken place. Therefore, participants' retrospective reports of perceived cohesion are construed here as *consequences* (i.e., outcomes) of the observed communication and performance metrics.

Framing the causal relationship between performance and communication, however, is especially more difficult for these data. On one hand, some tasks require communication between crew members, which has the implication that communication metrics should predict performance. On the other hand, observing one's team successes and failures will also motivate members to adjust their communication, implying that performance should predict communication. Without the time-series data to disentangle the temporal relationships between communication metrics and performance, there is apparently equal justification for either model. However, the communication metrics explored here (i.e., communication density and coherence) are thought to assess processes that contribute to the emergence of cohesion; therefore, communication metrics are treated here as *predictors* of performance, rather than outcomes. To summarize, the observed data were modeled on the following assumptions: 1) differences in communication cause changes in performance and 2) perceived cohesion, and 3) that differences in performance cause people to reflect differently on their reported levels of cohesion.

3. Results

All models were conducted using R version 3.6.2 (R Core Team 2019). Given the small sample size, these models should be considered primarily for descriptive purposes, although we do report standard errors and tests of statistical significance. Further along these lines, we use the reported p-values as to guide our discussion of patterns, which we feel are worth mentioning, and we used a very liberal guideline of interpreting coefficients with p-values at or below 0.20.

Due to the nested structure of these data, we used multilevel modeling using the *lme4* package (Bates et al. 2015) in R, and statistical tests were derived using normal approximation with the *lmerTest* package (Kuznetsova et al. 2017). Each of these models allowed for random variation in intercept estimates between dyads, but our sample size was not large enough for estimators to converge on a model with random slopes.

3.1 Descriptive Statistics

Due to the repeated measures, observations across conditions are not independent. To address this, repeated measures correlations (r_{rm}) were computed to represent the within-person relationships among variables across measurement occasions, which was accomplished using the *rmcorr* package (Backdash and Marusich 2020). Additionally, measurements were averaged across measurement occasions for each person and then correlated using a standard between-person Pearson's correlation coefficient (r). Both types of correlations for all outcomes are displayed in Table 3.

Table 3 Descriptive statistics for analyzed cohesion and performance variables

Variables	**Mean (SD)**	***Within-person*** **(r_{rm}; below diagonal) and between-person (r; above diagonal) correlations**						
		Belonging	**Morale**	**Density**	**Coher.**	**Mob. chal.**	**Resp. dist.**	**Threat prox.**
Belonging	5.74 (1.17)	…	**0.991*****	**–0.044**	**0.064**	**0.609+**	**–0.122**	**0.227**
Morale	5.81 (1.04)	*0.057*	…	**–0.154**	**0.043**	**0.615+**	**–0.206**	**0.321**
Density	1.59 (0.41)	*–0.262*	*0.122*	…	**–0.268**	**0.147**	**0.278**	**–0.929****
Coherence	0.105 (0.05)	*0.379+*	*–0.044*	*–0.654***	…	**–0.103**	**0.818***	**0.489**
Mob. Chal.	2.40 (0.66)	*–0.465+*	*–0.253*	*0.408+*	*–0.516**	…	**–0.271**	**0.118**
Resp. Dist.	23.92 (5.40)	*0.527**	*0.264*	*–0.069*	*0.032*	*–0.701***	…	**-0.097**
Threat Prox.	2.13 (1.20)	*–0.243*	*0.16*	*0.516**	*–0.077*	*0.455+*	*–0.457+*	…

Note: $+p < 0.20$, $*p < 0.05$, $**p < 0.01$, $**p < 0.001$.

There were notable differences between the within-person and between-person correlations, illustrating the need to account for the non-independence of responses over the sessions. Perhaps most striking are the correlations between the two self-reported cohesion factors, belongingness and morale, which bear no relationship within people over time ($r_{rm} = 0.057$), but correlate almost perfectly across people

(r = 0.991), suggesting that these constructs *within* people are operating in a fundamentally different way than these constructs *between* people. The data used to report these correlations were at the individual level, as opposed to the dyadic level, so they do not accurately represent the within-dyad and between-dyad relationships modeled next.

3.2 Analysis of Change in Cohesion and Communication Across Conditions

Because the same team interacted on multiple occasions, we examined the degrees to which self-reported cohesion, communication density, and dyadic coherence changed over time. To do so, we conducted four linear mixed effects models, which took the following form:

$$Level\ 1:$$

$$y_{ij} = \beta_0 + \beta_1\left(mobility_{ij}\right) + \beta_2\left(WMI_{ij}\right) + \beta_3\left(mobility * WMI_{ij}\right) + e_{ij}$$

$$Level\ 2:$$

$$\beta_0 = \gamma_{00} + u_{0j}$$

$$\beta_1 = \gamma_{10}$$

$$\beta_2 = \gamma_{20}$$

$$\beta_3 = \gamma_{30}$$

At Level 1, y_{ij} represents each outcome, β_0 represents the expected value of the outcome during autonomous planning (Mobility planning = 0) and Baseline WMI (WMI = 0), β_1 represents the expected difference in each outcome when comparing mobility planning conditions (Mobility planning: Autonomy = 0; Operator = 1), β_2 represents the expected difference in each outcome when comparing baseline and Transparent WMI conditions (Baseline = 0; Transparent = 1), β_3 represents the interaction between mobility and WMI conditions (Autonomy Planning-Baseline WMI vs. Operator Planning-Transparent WMI), and e_{ij} represents a normally distributed error term. At Level 2, u_{0j} represents the inclusion of a random intercepts term, which allows the estimate of the intercept, β_0, to vary across dyads.

3.2.1 Predicting Self-Reported Cohesion

Cohesion scores were computed separately for belongingness and morale, which were analyzed separately. Dyads' reported levels of belongingness caused difficulties for comparisons across conditions (Fig. 7)—specifically, all dyads reported an average of 6 on the 7-point scale when using Transparent WMI with operator planning, resulting in an invariant outcome for that condition. Similarly, with respect to morale, all dyads reported an average of 6 on the 7-point scale when using Baseline WMI with operator planning. Therefore, to examine the effects of experimental conditions on belongingness and morale, main effects models were used to compare mobility and WMI conditions, but their interactive effects are not considered (i.e., β_3, in the model described previously, was not estimated).

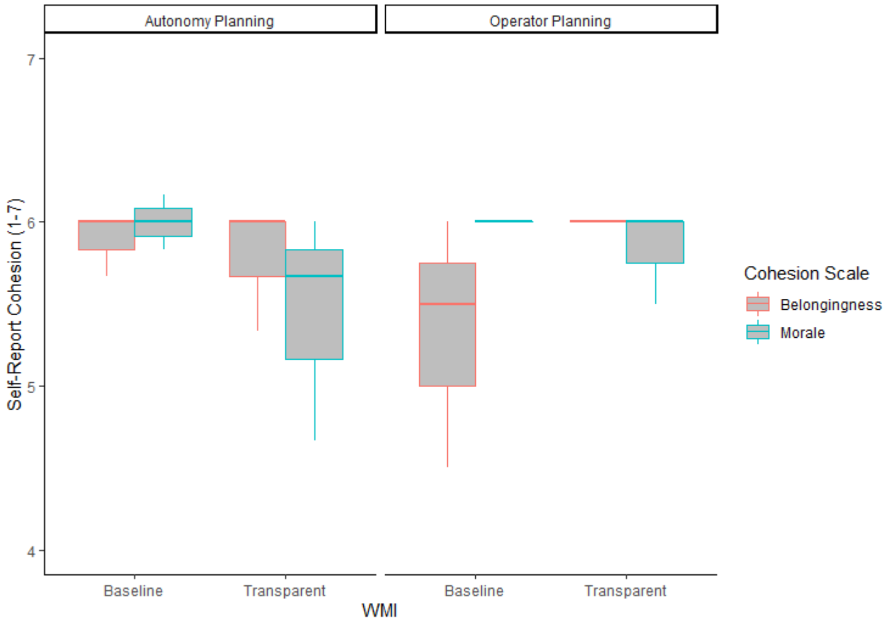

Fig. 7 Box plots showing the distribution of individual responses on the Perceived Cohesion Scale across conditions. Note that, during operator planning, there was no variability in responses in morale for Baseline WMI, nor was there any variability in belongingness.

Results indicated that operator planning, compared to autonomous planning, had no effect on self-reported ratings of belongingness, *b* = –0.167, se = 0.265, *t*(7) = –0.628, *p* = 0.55, nor did Transparent WMI improve self-reported ratings of belongingness above Baseline WMI, *b* = 0.278, se = 0.265, *t*(7) = 1.047, *p* = 0.33.

Although autonomous planning did not appear to impact self-reported ratings of morale, *b* = 0.194, se = 0.212, *t*(7) = 0.916, *p* = 0.390, it appeared that Transparent WMI, compared to baseline, may have negatively impacted dyads' reported levels of morale, *b* = –0.361, se = 0.212, *t*(7) = –1.701, *p* = 0.133.

Together, these results indicate that autonomous planning may not have any direct relationship with self-reported team cohesion, though WMI transparency may impact crew members' reported levels of morale, a critical dimension of team cohesion. However, due to the invariant responses in two conditions, there is no way to tell how robust these results would be in a larger sample.

3.2.2 Predicting Dyadic Communication Density

Communication density represents the rate of information flow across utterances, or the degree to which words relate to each utterance in semantic space—the more semantically relevant words for any given utterance would indicate that dyads are communicating more efficiently. Our model indicated that communication density did not differ between mobility planning conditions, *b* = 0.038, *se* = 0.151, *t*(6.91) = 0.255, *p* = 0.806; however, communication density appeared to be lower when using Transparent WMI compared to baseline, *b* = –0.383, *se* = 0.178, *t*(4.45) = –2.155, *p* = 0.091. Our results also suggest that the effects of WMI transparency may differ depending on mobility constraints, indicated by the positive interaction term, *b* = 0.296, se = 0.199, *t*(6) = 1.486, *p* = 0.188.

These results support the notion that transparency of autonomy-based decision making results in less-efficient verbal communication between crew members, perhaps because the autonomy's transparency makes dyads less dependent on this communication modality. If correct, this may indicate that dyads may have felt free to communicate less efficiently, because there was an increased shared situation awareness of the larger human–autonomy team. Further, if the interaction term is robust, it may indicate that having a Transparent WMI increases communication efficiency when planning must be carried out by the operator, rather than the autonomy.

3.2.3 Predicting Dyadic Coherence

Coherence provides a means to quantify if dyadic partners are speaking about similar topics. Our model indicated that coherence among dyads did not differ between mobility planning conditions, *b* = 0.031, *se* = 0.028, *t*(6.53) = 1.131, *p* = 0.298; however, dyadic coherence appeared to be greater when using Transparent WMI compared to baseline, *b* = 0.094, *se* = 0.029, *t*(6.50) = 3.242, *p* = 0.016. There was no indication that the effects of WMI transparency on dyadic coherence differed depending on mobility constraints, *b* = –0.053, se = 0.039, *t*(5.96) = –1.353, *p* = 0.2251. These results suggest that Transparent WMI may have facilitated more cohesive communication among team members. This could be viewed in contrast to the finding that communication density decreased with Transparent WMI. If Transparent WMI reduces the need for dyads to communicate regarding the task, perhaps communication efficiency decreases because dyads are speaking about topics unrelated to the task, which could have the effect of making dyads more engaged, and therefore speaking more similarly, in the task-irrelevant conversation.

3.3 Predicting Self-Reported Cohesion from Communication

To predict self-reported cohesion from dyadic communication density and dyadic coherence, four linear mixed effects models were constructed as follows:

$$Level\ 1:$$

$$Cohesion_{ij} = \beta_0 + \beta_1\left(Communication_{ij}\right) + e_{ij}$$

$$Level\ 2:$$

$$\beta_0 = \gamma_{00} + u_{0j}$$

$$\beta_1 = \gamma_{10}$$

where, at Level 1, $Cohesion_{ij}$ represents a dimension of self-reported cohesion (morale or belongingness), β_0 represents the expected value of cohesion when the communication metric (density or coherence) equals 0, β_1 represents the expected change in self-reported cohesion for every unit increase in the communication metric, and e_{ij} represents a normally distributed error term. At Level 2, u_{0j} represents the inclusion of a random intercepts term, which allows the estimate of the intercept, β_0, to vary across dyads.

When predicting self-reported belongingness, communication density did not have a noticeable influence, *b* = –0.207, se = 0.392, *t*(3.954) = –0.528, *p* = 0.626, but our results seemed to indicate that dyadic coherence may have a positive impact on perceived belongingness, *b* = 4.553, se = 2.553, *t*(10) = 1.783, *p* = 0.105. If this relationship is robust, it would indicate that team members' sense of belongingness is informed by the degree to which their utterances were semantically coherent with their partners' utterances.

When predicting self-reported morale, communication did not have a noticeable influence, *b* = 0.032, se = 0.332, *t*(2.906) = 0.095, *p* = 0.930, nor did dyadic coherence, *b* = –0.167, se = 2.582, *t*(9.492) = –0.065, *p* = 0.950.

Together, these results indicate that coherence between speakers' utterances may be informative for understanding team members' sense of belongingness, but not necessarily their sense of morale, whereas the density with which dyads communicate has no apparent impact on either dimension of cohesion. These findings also suggest that LSA-derived communication metrics are non-redundant with self-report measures, although these results do not provide much confidence that these metrics can be used interchangeably with constructs of belongingness and morale.

3.4 Predicting Cohesion from Performance

Ultimately, team cohesion is worth understanding insofar as it is characteristic of well-performing teams. Here, we analyze whether self-reported cohesion, at the dyadic level (i.e., the same level at which we assessed density and coherence in their communications), was associated with observed performance, the indicators of which are outlined in the methods section. To refresh, we analyzed the average time dyads spent faced with a *mobility challenge* (i.e., moving at rate less than 1 m/s for at least 5 s—more time is worse), the average *distance* from a threat when the dyad responded (further is better), and the *average proximity* of a dyad for each threat in the environment (i.e., at each time step during which a threat is in range, accumulate the distance from the threat—greater average proximity is worse).

To predict self-reported cohesion from performance metrics, six linear mixed effects models were constructed as follows:

$$Level\ 1:$$

$$Cohesion_{ij} = \beta_0 + \beta_1(Performance_{ij}) + e_{ij}$$

$$Level\ 2:$$

$$\beta_0 = \gamma_{00} + u_{0j}$$

$$\beta_1 = \gamma_{10}$$

where, at Level 1, $Cohesion_{ij}$ represents a dimension of self-reported cohesion (morale or belongingness), β_0 represents the expected value of cohesion when the performance metric (mobility challenge, response distance, or proximity to threat) equals 0, β_1 represents the expected change in self-reported cohesion for every unit increase in the performance metric, and e_{ij} represents a normally distributed error term. At Level 2, u_{0j} represents the inclusion of a random intercepts term, which allows the estimate of the intercept, β_0, to vary across dyads. As in the previous models, we do not allow for random slopes due to the limited sample size.

3.4.1 Mobility Challenges

Average time spent on mobility challenges had no impact on self-reported belongingness, *b* = –0.031, se = 0.169, *t*(8.795) = –0.184, *p* = 0.858, nor self-reported morale, *b* = 0.056, se = 0.150, *t*(8.962) = 0.373, *p* = 0.718, suggesting that difficulties with the autonomy did not inform participants' perceptions of cohesion.

3.4.2 Response Distance

Average response distance to threats did not affect self-reported belongingness, $b = 0.030$, se = 0.022, $t(10) = 1.339$, $p = 0.210$, nor self-reported morale, $b = 0.006$, se = 0.021, $t(10) = 0.275$, $p = 0.789$, suggesting that prompt, or delayed, responses to threats did not inform participants' perceptions of cohesion.

3.4.3 Threat Proximity

Average proximity to threats did not affect self-reported belongingness, $b = –0.035$, se = 0.119, $t(9.90) = –0.292$, $p = 0.777$. Interestingly, there did seem to be a positive effect of average threat proximity on morale, indicating that dyads appeared to report greater morale after experiencing more time in close proximity with threats, $b = 0.161$, se = 0.093, $t(10) = 1.729$, $p = 0.115$. On one hand, this latter result is counterintuitive, considering that better performance is thought to contribute to a greater sense of morale. On the other hand, increased risk to the crew may have been experienced as more exciting, a phenomenon known to increase social bonding (e.g., Aron and Aron [1996]), thereby increasing their self-report ratings of morale. However, we caution over-interpretation of this nonsignificant finding due to the small sample.

Overall, these results indicate that performance did not have much impact on participants' reported levels of cohesion.

3.5 Predicting Performance from Communication

To predict performance from communication metrics, three linear mixed effects models were constructed as follows:

$$Level\ 1:$$

$$Performance_{ij} = \beta_0 + \beta_1\left(Communication_{ij}\right) + e_{ij}$$

$$Level\ 2:$$

$$\beta_0 = \gamma_{00} + u_{0j}$$

$$\beta_1 = \gamma_{10}$$

where, at Level 1, $Performance_{ij}$ represents a performance metric (mobility challenge, response distance, or threat proximity), β_0 represents the expected performance value when communication metrics (density and coherence) equal 0, β_1 represents the expected change in performance for every unit increase in the communication metric (density or coherence), and e_{ij} represents a normally distributed error term. At Level 2, u_{0j} represents the inclusion of a random intercepts term, which allows the estimate of the intercept, β_0, to vary across dyads. As in the previous models, we do not allow for random slopes due to the limited sample size.

3.5.1 Predicting Mobility Challenges

Communication density appeared to contribute to more time spent facing mobility challenges, $b = 0.779$, se = 0.565, $t(10) = 1.379$, $p = 0.198$, whereas dyadic coherence appeared to contribute to less time spent facing mobility challenges, $b = –7.797$, se = 4.802, $t(10) = –1.624$, $p = 0.136$. If robust, these patterns indicate that dyads who speak more efficiently may be hindering each other as they navigate challenges with the autonomous system, whereas dyads communicating "on the same page" seems to bolster their abilities to navigate these challenges.

3.5.2 Predicting Response Distance

Communication density had no impact on the distance at which dyads responded to threats, $b = –1.135$, se = 4.483, $t(3) = –0.253$, $p = 0.817$, nor did dyadic coherence, $b = 28.496$, se = 37.091, $t(10) = 0.768$, $p = 0.460$. These results indicate that these communication metrics may not inform with how quickly (or, more accurately, how *closely*) participants responded to threats.

3.5.3 Predicting Average Threat Proximity

Communication density had no impact on how long dyads spent in close proximity to threats, $b = 1.216$, se = 1.163, $t(6.776) = 1.046$, $p = 0.332$, nor did dyadic coherence, $b = 0.389$, se = 7.722, $t(9.709) = 0.05$, $p = 0.961$. As with predicting response distance, these communication metrics have no apparent ability to help illuminate dyadic differences in how much time crews "linger" around threats.

Overall, the communication metrics explored here appear to have little to do with the focal performance metrics, with one exception—the degree to which dyads engaged in mobility challenges appeared to be due to dyads' communication density and coherence. Interestingly, these communication metrics had opposite effects, with greater density corresponding to more time spent facing challenges, but greater coherence corresponding to less time spent facing challenges.

4. Discussion

This report examined LSA as a tool for deriving metrics of team cohesion, using transcripts of team communications within a mixed human–autonomy crew. Specifically, LSA was used to derive metrics of communication density, which represents a tendency to use fewer words to convey more information, and between-speaker coherence, which represents the tendency for speakers to discuss similar topics. These metrics were then used to predict dimensions of cohesion (from self-report data) and performance (from behavioral data). Results indicated that communication density may be positively associated with task challenges, meaning that greater communication density may have a negative impact on team dynamics. However, results also indicated that coherence between speakers may predict greater perceived belongingness and less time spent overcoming task-related challenges, meaning that coherence between teammates appears to have a positive impact on team dynamics.

These findings also had important implications for understanding dynamics in a human–autonomy team. For instance, greater transparency in the autonomy's decision-making reduced the efficiency with which dyads communicated, while the transparency increased the coherence of the dyadic communication. Together, these results may be indicative of less formal but more engaged conversations between crew members when the burden of task-relevant communication is offloaded by the Transparent WMI. Given the observed positive relationship between dyadic coherence and belongingness, these results may indicate that transparent autonomous systems increase cohesion by making crew members more *interpersonally* engaged. Further, in terms of performance, the observed negative relationship between dyadic coherence and time spent facing mobility challenges may indicate that this interpersonal engagement does not come with a detriment to performance; indeed, if the crew found it easier to perform with a transparent autonomous system, they may have felt even freer to engage interpersonally.

To prevent taking these speculative claims too far, it is crucial to acknowledge the limitations of the data analyzed here, the most obvious of which is the sample size. Although the repeated measures nature of these data enabled meaningful comparisons across experimental conditions, the fact that only one crew was analyzed makes the generalizability of our claims highly speculative. Beyond the limited information one can glean from any analysis of these data, the analyses themselves were also highly restricted. For instance, with a multilevel framework, in which multiple teams would be measured on multiple occasions, it is recommended that predictive constructs are structured in a way that disentangles within-cluster (with "cluster" being the individual, dyad, or crew measured over multiple occasions) from between-cluster effects (Enders and Tofighi 2007). If this could be done under the framework explored here, one could make claims, for instance, about the effects of *dyads who exhibit higher coherence* and, separately, about the effects of *when dyads exhibit higher coherence.* This was addressed, in part, by separating the within-individual correlations from the between-individual correlations in Table 3, which revealed some interesting divergences between interindividual and intra-individual correlations. For example, when people communicated with greater density, they tended to remain closer to threats for longer (r = 0.516), but people who tended to communicate with greater density overall spent less time in close proximity to threats (r = –0.929). However, the between-*individual* estimates, computed with six data points, would not reflect the between-*dyad* estimates, computed with just three data points. Therefore, if this work is to inform future interventions for improving team cohesion and performance, enough data from enough teams would need to be collected to truly disentangle the levels at which these constructs are impacting cohesion and performance.

This report is also limited due to the manner in which these communication data were collected and transcribed. Future technologies will require an automated transcription process with enough accuracy to produce valid, real-time communications analyses. Fortunately, LSA appears to be robust to some level of transcription inaccuracy, as one report found just a 10% decrease in performance of LSA conducted on a transcription with just 57% accuracy (Foltz et al. 2006). This study was equipped to handle transcriptions using automated speech-to-text software, but issues with audio quality prevented the automated process from reaching a tolerable degree of accuracy. Therefore, analyses were limited to those communications that were manually transcribed.

Finally, this report is also limited due to several relevant variables that were not collected here, but would be critical for drawing conclusions in future analyses. Although there exist several individual and team constructs that would be informative for examining nuanced team dynamics, two constructs would have been particularly informative for this novel research area: perceived communication and task cohesion. First, perceived communication metrics, such as communication quality, would provide an excellent check against the LSA-derived communication metrics. Second, task cohesion would provide a more relevant measure of team cohesiveness than the interpersonal cohesion captured by belongingness and morale. To this point, the results reported here indicate a tenuous link between communication metrics and self-reported metrics of cohesion, meaning that these LSA-derived communication metrics are not indicative of team cohesion, that these metrics tap into features of cohesion that cannot be characterized by belongingness and morale, or that our sample was too small to detect any real associations. Therefore, measuring the effects of communication metrics on other team cohesion constructs would be a logical next step toward understanding how communication affects human–autonomy team dynamics.

5. Conclusions

Traditionally, teams' researchers have examined cohesion using standard psychometric measures, with the degree of team cohesion often representing some aggregate of team members' reported levels of cohesion. Although such methods have yielded important insights over decades of research, future technologies will require deriving these higher-level constructs from low-level behavioral data. This is particularly true for constructs such as team cohesion, which is known to emerge from the lower-level interpersonal decisions made by team members—after all, depending on self-report measures requires retrospective summary accounts, which gloss over the dynamic nature of the interactions, or would require frequent interruptions that divert members from their tasks and goals.

6. References

Abrams AM, Rosenthal-von der Pütten AM. I–C–E framework: concepts for group dynamics research in human-robot interaction. International Journal of Social Robotics. 2020;1–17.

Ahronson A, Cameron JE. The nature and consequences of group cohesion in a military sample. Military Psychology. 2007;19:9–25.

Aron EN, Aron A. Love and expansion of the self: the state of the model. Personal Relationships. 1996;3(1):45–58.

Babcock MJ, Ta VP, Ickes W. Latent semantic similarity and language style matching in initial dyadic interactions. Journal of Language and Social Psychology. 2014;33(1):78–88.

Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. Journal of Statistical Software. 2015;67(1):1–48. doi:10.18637/jss.v067.i01.

Beal DJ, Cohen RR, Burke MJ, McLendon CL. Cohesion and performance in groups: a meta-analytic clarification of construct relations. Journal of Applied Psychology. 2003;88(6):989.

Behoora I, Tucker CS. Machine learning classification of design team members' body language patterns for real time emotional state detection. Design Studies. 2015;39:100–127.

Borsboom D. When does measurement invariance matter? Medical Care. 2006;44(11):S176–S181.

Casey-Campbell M, Martens ML. Sticking it all together: a critical assessment of the group cohesion-performance literature. International Journal of Management Reviews. 2009;11:223–246.

Chhan D, Scharine A, Perelman BS. Human–autonomy teaming essential research program project 2: Transparent multimodal crew interface designs. Technical Note 3: Multimodal cueing for transparency in mobility operations. Aberdeen Proving Ground (MD): CCDC Army Research Laboratory (US); 2020. Report No.: ARL-TN-1019.

Curtis RC, Miller K. Believing another likes or dislikes you: behaviors making the beliefs come true. Journal of Personality and Social Psychology. 1986;51:284–290.

DeChurch LA, Mesmer-Magnus JR. The cognitive underpinnings of effective teamwork: a meta-analysis. Journal of Applied Psychology. 2010;95(1):32.

Dhall A. EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks. In: 2019 International Conference on Multimodal Interaction; 2019 Oct. p. 546–550.

Dobbins GH, Zaccaro SJ. The effects of group cohesion and leader behavior on subordinate satisfaction. Group and Organization Studies. 1986;11:203–219.

Dong A. The latent semantic approach to studying design team communication. Design Studies. 2005;26(5):445–461.

Dong A, Hill AW, Agogino AM. A document analysis method for characterizing design team performance. J Mech Des. 2004;126(3):378–385.

Dumais ST. Latent semantic analysis. Annual Review of Information Science and Technology. 2004;38(1):188–230.

Ehsan N, Mirza E, Ahmad M. Impact of computer-mediated communication on virtual teams' performance: an empirical study. Proc 2008 International Symposium on Information Technology. 2008 Aug;3:1–8. IEEE.

Enders CK, Tofighi D. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. Psychological Methods. 2007;12(2):121.

Foltz PW, Martin MJ. Automated communication analysis of teams. In: Salas E, Goodwin G, Burke CS, editors. Team effectiveness in complex organizations. Abingdon (UK): Routledge; 2008. p. 411–432.

Foltz PW, Martin MJ, Abdelali A, Rosenstein MB, Oberbreckling RJ. Automated team discourse modeling: test of performance and generalization. In Proceedings of the 28th Annual Cognitive Science Conference; 2006 July. p. 1317–1322.

Gigone D, Hastie R. The common knowledge effect: Information sharing and group judgment. Journal of Personality and Social Psychology. 1993;65(5):959.

Gorman JC, Foltz PW, Kiekel PA, Martin MJ, Cooke NJ. Evaluation of latent semantic analysis-based measures of team communications content. Proc Human Factors and Ergonomics Society Annual Meeting. 2003 Oct;47(3):424–428. Los Angeles (CA): SAGE Publications.

Gorman JC, Martin MJ, Dunbar TA, Stevens RH, Galloway T. Analysis of semantic content and its relation to team neurophysiology during submarine crew training. In: International Conference on Augmented Cognition; 2013 July. p. 143–152. Springer, Berlin/Heidelberg, Germany.

Grossman R. How do teams become cohesive. A meta-analysis of cohesion's antecedents [doctoral dissertation]. [Orlando (FL): University of Central Florida]; 2014.

Gruenfeld DH, Mannix EA, Williams KY, Neale M.A. Group composition and decision making: how member familiarity and information distribution affect process and performance. Organizational Behavior and Human Decision Processes. 1996;67(1):1–15

Hill AW, Dong A, Agogino AM. Towards computational tools for supporting the reflective team. In: Artificial Intelligence in Design'02; 2002. p. 305–325. Springer, Dordrecht, Germany.

James LR, Demaree RG, Wolf G. Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology. 1984;69(1):85.

James LR, Demaree RG, Wolf G. rwg: An assessment of within-group interrater agreement. Journal of Applied Psychology. 1993;78(2):306–309. https://doi.org/10.1037/0021-9010.78.2.306.

Janis IL. Groupthink: Psychological studies of policy decisions and fiascoes. Boston (MA): Cengage Learning; 1982.

Kiekel PA, Cooke NJ, Foltz PW, Gorman JC, Martin MJ. Some promising results of communication-based automatic measures of team cognition. Proc Human Factors and Ergonomics Society Annual Meeting. 2002;46. doi: 10.1177/154193120204600318.

Kim S, Magnusen MJ, Andrew DP. Divided we fall: examining the relationship between horizontal communication and team commitment via team cohesion. International Journal of Sports Science and Coaching. 2016;11(5):625–636.

Kozlowski SWJ. Advancing research on team process dynamics. Organizational Psychology Review. 2015;5(4):270–299. doi:10.1177/2041386614533586.

Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. Journal of Statistical Software. 2017;82(13):1–26. doi: 10.18637/jss.v082.i13.

Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes. 1998;25(2-3):259–284.

Liang HY, Shih HA, Chiang YH. Team diversity and team helping behavior: the mediating roles of team cooperation and team cohesion. European Management Journal. 2015;33(1):48–59.

Mach M, Dolan S, Tzafrir S. The differential effect of team members' trust on team performance: the mediation role of team cohesion. Journal of Occupational and Organizational Psychology. 2010;83(3):771–794.

Marks MA, Mathieu JE, Zaccaro SJ. A temporally based framework and taxonomy of team processes. Academy of Management Review. 2001;26(3):356–376.

Martin M, Foltz P. Automated team discourse annotation and performance prediction using LSA. In: Proceedings of HLT-NAACL 2004: Short Papers; 2004. p. 97–100.

Mathieu JE, Heffner TS, Goodwin GF, Salas E, Cannon-Bowers JA. The influence of shared mental models on team process and performance. Journal of Applied Psychology. 2000;85(2):273.

Mathieu J, Maynard MT, Rapp T, Gilson L. Team effectiveness 1997–2007: a review of recent advancements and a glimpse into the future. Journal of Management. 2008;34(3):410–476.

MIL-STD 2525D. Common warfare symbology (version 1). Washington (DC): Department of Defense (US); 1994 Sep 30. https://www.jcs.mil/Portals/36/Documents/Doctrine/Other_Pubs/ms_2525d.pdf.

Neubauer C, Woolley J, Khooshabeh P, Scherer S. Getting to know you: a multimodal investigation of team behavior and resilience to stress. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction; 2016 Oct. p. 193–200.

Neubauer C, Chollet M, Mozgai S, Dennison M, Khooshabeh P, Scherer S. The relationship between task-induced stress, vocal changes, and physiological state during a dyadic team task. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction; 2017 Nov. p. 426–432.

Ng KY, Van Dyne L. Antecedents and performance consequences of helping behavior in work groups: a multilevel analysis. Group and Organization Management. 2005;30(5):514–540.

Oliver LW, Harman J, Hoover E, Hayes SM, Pandhi NA. A quantitative integration of the military cohesion literature. Military Psychology. 1999;11(1):57-83.

Onağ Z, Tepeci M. Team effectiveness in sport teams: the effects of team cohesion, intra team communication and team norms on team member satisfaction and intent to remain. Procedia-Social and Behavioral Sciences. 2014;150:420–428.

Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001. LIWC2001. 2001;71.

Perelman BS, Lakhmani S, Wright JL. Human–autonomy teaming essential research program project 2: transparent multimodal crew interface designs. Technical note 1: general overview. Aberdeen Proving Ground (MD): CCDC Army Research Laboratory (US); 2020a. Report No.: ARL-TN-1003.

Perelman BS, Lakhmani S, Wright JL, Chhan D, Scharine A, Evans III AW, Gaston JR, Hill SG, Mermagen T, Ericson M, Hansberger JT, Roy H, Cox KR, Marathe AR. Human–autonomy teaming essential research program project 2: transparent multimodal crew interface designs. Project summary technical report. Aberdeen Proving Ground (MD): CCDC Army Research Laboratory (US); 2020b. Report No.: ARL-TR-9002.

R Core Team. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2019. https://www.R-project.org/.

Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation. 2004;60(5):503–520.

Santoro JM, Dixon AJ, Chang CH, Kozlowski SW. The dynamics of team cohesion: Methods, emerging tools, and advanced technologies. In: Salas E, Vessey W, Estrada A, editors. Team cohesion: advances in psychological theory, methods and practice. Research on Managing Groups and Teams. 2015;17:115–145.

Silva C, Ribeiro B. The importance of stop word removal on recall values in text categorization. Proc International Joint Conference on Neural Networks. 2003 July;3:1661–1666. IEEE.

Sperber D, Wilson D. Relevance: communication and cognition (vol. 142). Cambridge (MA): Harvard University Press; 1986.

TRADOC. The US Army Multi-Domain Operations, 2028. Ft. Eustis (VA): Training and Doctrine Command; 2018 Dec 8. https://www.tradoc.army.mil/Portals/14/Documents/MDO/TP525-3-1_30Nov2018.pdf.

Whitman LE, Malzahn DE, Chaparro BS, Russell M, Langrall R, Mohler BA. A comparison of group processes, performance, and satisfaction in face-to-face versus computer-mediated engineering student design teams. Journal of Engineering Education. 2005;94(3):327–337.

Wild F. lsa: Latent Semantic Analysis. R package version 0.73.2. R-Forge; 2020.

Woehr DJ, Loignon AC, Schmidt PB, Loughry ML, Ohland MW. Justifying aggregation with consensus-based constructs: a review and examination of cutoff values for common aggregation indices. Organizational Research Methods. 2015;18(4):704–737.

Appendix. Experimental Conditions by Day

	Day 1	**Day 2**	**Day 3**	**Day 4**	**Day 5**
AM (9-1pm)	• Welcome to lab • Consent form • Equipped with physio and baseline • Eye-trackers and calibration • MSSQ • Demographics • SSQ • Baseline 5 min TRAINING Run • SSQ • Removal of sensors	• Physio • Eye-trackers • SSQ • Baseline 5min CONSTRAINED BASELINE • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale • Removal of sensors	• Physio • Eye-trackers • SSQ • Baseline 5min OPEN P2 • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale • Removal of sensors	• Physio • Eye-trackers • SSQ • Baseline 5min OPEN BASELINE • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale • Removal of sensors	• Physio • Eye-trackers • SSQ • Baseline 5min CONSTRAINED P2 • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale • Removal of sensors
Lunch 1 1-1:30	all	Tablet Evals Group 1	all	all	all
Lunch 2 1:30-2:00		Tablet Evals Group 2			
PM (2-6)	• Physio • Eye-trackers • SSQ • Baseline 5min ~~RAILS BASELINE~~ (Training) • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale • Removal of sensors	• Physio • Eye-trackers • SSQ • Baseline 5min OPEN BASELINE • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale • Removal of sensors	• Physio • Eye-trackers • SSQ • Baseline 5min CONSTRAINED P2 • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale • Removal of sensors	• Physio • Eye-trackers • SSQ • Baseline 5min OPEN P2 • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale • Removal of sensors	• Physio • Eye-trackers • SSQ • Baseline 5min CONSTRAINED BASELINE • SSQ • SUS • Trust in Automation Scale • Team Trust Scale • Perceived Cohesion Scale Removal of sensors

List of Symbols, Abbreviations, and Acronyms

AFC	US Army Futures Command
ANVEL	Autonomous Navigation Virtual Environment Laboratory
ARL	Army Research Laboratory
CCDC	US Army Combat Capabilities Development Command
CFTs	Cross-Functional Teams
COTS	commercial off-the-shelf
HAT ERP	Human–Autonomy Teaming Essential Research Program
LSA	latent semantic analysis
MCV	manned combat vehicle
MDO	Multi-Domain Operations
NGCV	Next-Generation Combat Vehicle
RCV	robotic combat vehicle
SOMTE	Soldier, Operator, Maintainer, Test, and Evaluation
WMI	Warfighter Machine Interface

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 CCDC ARL
(PDF) FCDD RLD DCI
TECH LIB

1 CCDC ARL
(PDF) FCDD RLH B
T DAVIS
BLDG 5400 RM C242
REDSTONE ARSENAL AL
35898-7290

1 CCDC ARL
(PDF) FCDD HSI
J THOMAS
6662 GUNNER CIRCLE
ABERDEEN PROVING
GROUND MD
21005-5201

1 USAF 711 HPW
(PDF) 711 HPW/RH K GEISS
2698 G ST BLDG 190
WRIGHT PATTERSON AFB OH
45433-7604

1 USN ONR
(PDF) ONR CODE 341 J TANGNEY
875 N RANDOLPH STREET
BLDG 87
ARLINGTON VA 22203-1986

1 USA NSRDEC
(PDF) RDNS D D TAMILIO
10 GENERAL GREENE AVE
NATICK MA 01760-2642

1 OSD OUSD ATL
(PDF) HPTANDB B PETRO
4800 MARK CENTER DRIVE
SUITE 17E08
ALEXANDRIA VA 22350

ABERDEEN PROVING GROUND

18 CCDC ARL
(PDF) FCDD RLH
J LANE
Y CHEN
P FRANASZCZUK
A MARATHE
K MCDOWELL
K OIE
K SCHAEFER
B PERELMAN
FCDD RLH F
J GASTON (A)
FCDD RLH FA
A DECOSTANZA
C NEUBAUER
D FORSTER
FCDD RLH FB
D BOOTHE (A)
S MCGHEE
FCDD RLH FC
K COX (A)
FCDD RLH FD
A FOOTS (A)
A KRAUSMAN
FCDD RLH FE
D HEADLEY